A learning-based approach to text image retrieval: using CNN features and improved similarity metrics

Mao Tan, Si-ping Yuan, and Yong-xin Su*

Key Laboratory of Intelligent Computing & Information Processing
of Ministry of Education,
Xiangtan University, Xiangtan, 411105 China

Abstract. Text content can have different visual presentation ways with roughly similar characters. While conventional text image retrieval depends on complex model of OCR-based text recognition and text similarity detection, this paper proposes a new learning-based approach to text image retrieval with the purpose of finding out the original or similar text through a query text image. Firstly, features of text images are extracted by the CNN network to obtain the deep visual representations. Then, the dimension of CNN features is reduced by PCA method to improve the efficiency of similarity detection. Based on that, an improved similarity metrics with article theme relevance filtering is proposed to improve the retrieval accuracy. In experimental procedure, we collect a group of academic papers both including English and Chinese as the text database, and cut them into pieces of text image. A text image with changed text content is used as the query image, experimental results show that the proposed approach has good ability to retrieve the original text content.

Keywords: Text retrieval; Text image retrieval; Image Similarity; Convolutional neural networks; PCA

1 Introduction

With booming development of digital media technology, the scale of multimedia resources including the text images is getting bigger and bigger. Since text retrieval is a well studied problem in natural language processing, many approaches based on Optical Character Recognition (OCR) applications have been proposed, which recognize text content from images and then use text retrieval technologies to implement text image retrieval system.

While conventional text image retrieval depends on complex model of OCR-based text recognition and text similarity detection, direct recommendation and retrieval on the basis of arbitrary multi-character text in unconstrained image require a similarity retrieval approach to learn and recognize deep visual features in images. The new text image similarity retrieval approach will be conducive to detect the re-contributed and re-published text content on the database of academic journals theses, or query the relevant literature in massive resources.

In early studies on text recognition and retrieval, the extraction of features requires layout analysis, line segmentation, word segmentation, word recognition, etc. But over the last decade, deep learning based features extraction has become an key research direction. Among various deep learning models, the Convolutional Neural Networks (CNNs) are the most powerful networks in image processing tasks. When CNNs are trained on object recognition, a deep representation of the image is constructed to make object information increasingly explicit along the processing hierarchy [1]. During the CNN feature training phase, Redmon et al. [2] proposed a improved model that inspired by the GoogLeNet model [3] for image classification. They pre-trained the model's convolutional layers on ImageNet dataset for approximately a week, and used the initial convolutional layers of the network to extract features from the image while the fully connected layers to predict the result. Gatys et al. [4] obtained a style representation of an input image and generated results on the basis of the VGGNet, which is a CNN that rivals human performance on a common visual object recognition benchmark task [5]. As for recognizing multi-character text, Ian et al. [6] proposed a unified approach that integrates the localization, segmentation, and recognition steps via the use of a deep convolutional neural network that operates directly on the image pixels.

In the 2014 ImageNet ILSVRC competition, VGGNet secured the first and the second places in the localisation and classification tracks respectively. It increases depth using an architecture with very small (3×3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers [5]. However, compared with other similar methods, the parameter space of VGGNet is too large to train a VGGNet model in a short time. Fortunately, there are some open pre-trained model that we can easily use, such as MatConvNet [7]. Besides that, training the CNN features on a large dataset and fine-tuning by target dataset can significantly improve the performance [8]. Furthermore, we can use the PCA method to reduce the dimension of the CNN features according to the investigation in reference [9], which is mainly to evaluate the performance of compressed neural codes, it is declared that plain PCA or a combination of PCA with discriminative dimensionality reduction can result in very short codes and good (state-of-the-art) performance.

Similarity measurement is another key technique to determine the effectiveness of the retrieval system. There are many ways to measure the similarity of image content according to different attributes. In most image retrieval system, the similarity measurement between the query image and the image database is always employed to find out the original or similar image. Sejal et al. [10] computed keyword relevance between annotated keywords of images using absorbing Markov chain, and then images were ranked by keyword relevance probability for recommendation. In addition, a more efficient and widespread method is computing pair-wise image cosine similarity based on visual features of all images, then used this parameter value to retrieve the high similarity images [11]. Nevertheless, improving similarity retrieval precision is usually a difficulty in

practice if only use text image cosine similarity, i.e., different text images that have similar visual feature but different high-level theme features according to the association between the words.

In this paper, we try to establish a new learning-based approach to text image retrieval with the purpose of finding out the original or similar text through a query text image. According to different article themes, a text image similarity retrieval method with CNN feature extraction and cosine similarity matching is chosen as a basic framework, and a theme relevance filtering model is proposed and integrated to the framework to improve the accuracy of retrieval system. In the experimental procedure, we slice a batch of English and Chinese academic papers into text images as the image dataset, a text image with changed text content is used as the query image, several case studies are provided to evaluate the adaptability and accuracy of the proposed method for different conditions.

2 Methodology

In this section, we mainly discuss several key steps of the text image similarity retrieval based on the CNN features of images. Firstly, we use a pre-trained VGGNet model to extract CNN features from the experimental text image dataset, which can convert the visual content into a deep representation. As the CNN feature matrix trained by the VGGNet are high-dimensional, we further perform the PCA method to reduce the dimensions. Then we compute the cosine similarity of each pair of text images based on the CNN features. Besides that, the text image to be retrieved can reflect its relevance to a certain article themes by computing the similarity of text content, therefore, in order to further improve the accuracy of similarity retrieval, we propose an article theme relevance filtering model to adjustment the original similarity. In the following section, we elaborate on each of steps in detail, the entire processes are shown in Figure 1.

2.1 CNN Training with VGGNet

It is necessary to extract the primitive features of text image as the constructive parameter of the training model. The quality of the feature extraction directly determine the retrieval effect. Therefore, we need to use image processing technologies and mathematical methods to extract the appropriate characteristics information in the visual content. Recently, CNNs have achieved impressive results in some areas such as image recognition and object detection. It can input image into the network directly, avoiding the complex feature extraction and data reconstruction process in traditional recognition algorithm. In addition, VGGNet model is a preferred multi-layer neural network model for extracting the CNN features of image. The VGGNet use small-size convolution filters and deep network layers, which has strong generalization ability in many different computer vision applications and other image recognition datasets. We choose the excellent VGGNet-E network that introduced in [5] as the training model, which comprises 19 learnable layers, the previous 16 are convolutional layers,

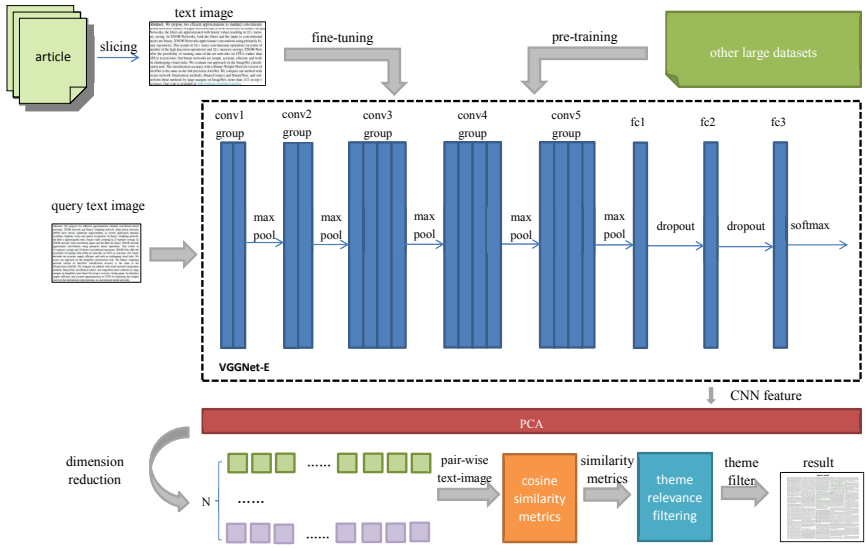

Fig. 1. The entire process of similar arbitrary multi-character text image retrieval. **CNN Training with VGGNet.** Firstly, extract the CNN feature to obtain the deep visual representations by fine-tuning the pre-trained VGGNet model on the target text image dataset. **Dimension Reduction by PCA.** After obtaining the CNN feature matrix, reduce the dimension of the matrix by PCA to improve the efficiency of the algorithm. **Similarity Calculation.** Measure the cosine similarity between the query text image and each image in the dataset based on the low-dimensional CNN features. **Theme Relevance Filtering.** Finally, statistic the degree of relevance between the query text image and the various article theme to improve the similarity metrics.

and the last 3 are fully-connected layers. The structure and parameters of each layer in VGGNet-E model are shown in Table 1.

The detail parameters of each convolutional layer are listed in three sub-columns of Table 1. The first column specifies the number of convolution filters and their receptive field size as 'num $\times$ size $\times$ size'; the second column indicates the convolution stride ('st.') and spatial padding ('pad'); the third column indicates the max-pooling down sampling factor. For fully-connected layers, the first sub-column specify their dimensionality. FC17 and FC18 use dropout method to be regularised and the last layer FC19 acts as a multi-way soft-max classifier. Among them, the activation function of VGGNet-E is the rectification Linear Unit (RELU). The CNN training through VGGNet-E model is carried out by using mini-batch gradient descent. The batch size, momentum, the L2 penalty multiplier of weight decay and dropout regularisation ratio for the fully-connected layers are set to 256, 0.9, 0.0005, 0.5 respectively. The learning rate is set to 0.01 initially, and will decrease to one-tenth of the initial value when the validation set accuracy improving is stopped.

Table 1. VGGNet-E model's structure and parameters.

Layers	Conv filters' number & Receptive field size	Conv stride & Spatial padding	Down sampling
Conv1	64×3×3	st. 1, pad 1	–
Conv2	64×3×3	st. 1, pad 1	x2 pool
Conv3	128×3×3	st. 1, pad 1	–
Conv4	128×3×3	st. 1, pad 1	x2 pool
Conv5	256×3×3	st. 1, pad 1	–
Conv6	256×3×3	st. 1, pad 1	–
Conv7	256×3×3	st. 1, pad 1	–
Conv8	256×3×3	st. 1, pad 1	x2 pool
Conv9	512×3×3	st. 1, pad 1	–
Conv10	512×3×3	st. 1, pad 1	–
Conv11	512×3×3	st. 1, pad 1	–
Conv12	512×3×3	st. 1, pad 1	x2 pool
Conv13	512×3×3	st. 1, pad 1	–
Conv14	512×3×3	st. 1, pad 1	–
Conv15	512×3×3	st. 1, pad 1	–
Conv16	512×3×3	st. 1, pad 1	–
Layers	Dimensionality	Operation	
FC17	4096×7×7	dropout	
FC18	4096×1×1	dropout	
FC19	1000×1×1	softmax	

In general, most CNNs model are trained by composing simple linear and non-linear filtering operations, while their implementation need to be trained on large dataset and learned from vast amounts of data, usually millions of images. Therefore, we fine-tune a state-of-the-art pre-trained VGGNet model on the target text image dataset. At the training phase, we scale down the original text image proportionally, ensuring that the smallest side of this isotropically rescaled training image is greater than 224. And then, we crop the training image randomly and the image size is fixed to 224×224. Through a series of preprocessing, this network model takes text images of a fixed size as input and removes the mean. After that, we retain the CNN feature matrix of the penultimate layer of this deep CNN representation, which can be used as a powerful image descriptor applicable to many types of datasets.

2.2 Dimension Reduction by PCA

After the CNN feature extraction with VGGNet-E model, we obtain a 4096-D image deep representation, which is often too high to compute the complex feature similarity, resulting in excessive consumption of memory computing resources. Therefore, it is necessary to adopt a method to reduce feature dimensionality. In addition, experiments in [9] show that the performance of different versions of neural codes after PCA compression to a different number of dimensions works surprisingly well, thus we use the PCA method to compress the CNN feature matrix to 128-D almost without any loss of the retrieval accuracy. It maps the high-dimensional features through a linear transformation to a low-dimensional space to reduce some information redundancy, and reflects the original text image more effectively.

One of the main ideas of PCA is to discover the hidden feature information in the internal structure of high-dimensional deep representation, and exclude the irrelevant information. Its purpose is to retain the hidden structure that can capture most of the original features, and avoid the trained model learning from a large number of noise data. In order to avoid the influence of the sample units, and simplify the calculation of covariance matrix, we process the CNN feature matrix by removing mean value normalization method, and then use the PCA method to find the 128 largest variation feature vectors in this matrix.

Therefore, covariance matrix C can be calculated according to each feature vector X_i in normalized CNN feature matrix, which can be expressed as

$$C = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T, \tag{1}$$

where C represents the covariance matrix of the feature matrix, and n represents the number of feature vectors, that is, 4096.

After that, the eigenvalue equation based on C can be expressed as

$$\lambda_i \mu_i = C \mu_i, \tag{2}$$

where λ_i is the eigenvalue of the covariance matrix, and μ_i is the corresponding eigenvector of the covariance matrix.

Then, we use the resulting 128 normalized feature vectors to constitute the main feature matrix to form a 128-D space. Based on that, we project the 4096-D dimensional CNN feature matrix onto the 128-D dimensional space. Finally, the CNN feature projection matrix is indexed to improve the retrieval efficiency.

2.3 Similarity Calculation

Cosine similarity has been proved to be an effective metric system because of its accuracy. The 128-D feature matrix$[z_1, z_2, \ldots, z_n]^T$ describe the main CNN features of the text image in the dataset, where n is the number of text images in the datasets. The cosine similarity calculated from each pair of text image's CNN feature vector can approximately measure the similarity between text images. Cosine similarity framework uses the cosine of two vectors' angle to measure the difference between two text images, it focuses on the direction difference of the vector.

For each pair of text images' feature vector (Z_u, Z_v) where $u \neq v$, the pair-wise text image cosine similarity T_s can be expressed as

$$T_s(Z_u, Z_v) = \frac{\sum_{i=1}^{K} F(Z_u, u_i) * F(Z_v, v_i)}{\sqrt{\sum_{i=1}^{K} F(Z_u, u_i)^2 * \sum_{i=1}^{K} F(Z_v, v_i)^2}}, \tag{3}$$

where $K = 128$, and $F(Z_u, u_i)$ is the value of the i-th column element of the 128-D dimensional feature vector corresponding to the text image Z_u. $T_s(Z_u, Z_v)$ is the pair-wise text image cosine similarity. The denominator is the vector length named euclidean distance, which can be expressed as a L2-norm when the CNN feature matrix is L2 normalized. Thus, the pair-wise text image cosine similarity T_s can also be calculated as

$$T_s(Z_u, Z_v) = \frac{Z_u \bullet Z_V}{\|Z_u\|_2 \|Z_v\|_2}. \tag{4}$$

Through the calculated similarity of the pair-wise text image, we can retrieve out some high similarity text images. The similarity's value is between -1 and 1, where the value is 1 means that the pair-wise text images are completely similar, and 0 represents they are not related, even have no similarity. At the same time, cosine similarity can also capture the negative correlation. When the similarity's value is -1, it means not only that the two text images' content are not related, but also they are completely different.

2.4 Theme Relevance Filtering

For any given query text image, statistical analysis shows that the cosine similarity value related to certain specific article themes will always be higher than others in the retrieval results. which causes this may be the small size filter used

in VGGNet that captures the basic pattern to extract low-level features, and deeper network that combines low-level features into high-level features, which directly reflect the relevance among text content. However, in practical applications, some text images from different themes article is more approximate, because of the visual features extracted from a short text content in the images are limited inherently, as well as most of their local visual features may be similar. Therefore, we design an article theme relevance filtering model according to the theme relevance degree of the query text image. Firstly, we calculate the each article theme similarity whose value depend on the feedback of its cosine similarity. Then we use the article theme similarity to set the weighting function and reorder the preliminary result.

In the preliminary result, we figured out the cosine similarity between the query text image p^* and each text image in the dataset, which are arranged by similarity in descending order. Firstly, we index the text images by the file names, which reflects the article themes. After that, the article theme similarity mean L_i to each theme can be calculated as

$$L_i = \frac{\sum_{j=1}^{M} T_s(Z_{p^*}, {Z_{L_i}}^{(j)})}{M}, \tag{5}$$

where L_i denotes the article theme similarity mean of the i-th article themes, and ${Z_{L_i}}^{(j)}$ represents the cosine similarity between the query text image and the j-th text image that correspond to the i-th article themes, the total images correspond to the i-th article themes in the preliminary results is M.

Based on this, we construct an theme relevance filtering model, in which the weights of article theme similarity are set by the article theme similarity mean, and the preliminary results are reordered by the weights in descending order. Thus, the improved similarity metric T_s' can be calculated as

$$T_s'(Z_{p^*}, Z_{L_i}) = T_s(Z_{p^*}, Z_{L_i}) * L_i. \tag{6}$$

3 Experiments

3.1 Data Collection

In this work, in order to carry out the text image similarity retrieval tests on different article theme, we collect a group of English and Chinese academic papers as the text database, and cut them into many small pieces of text image to construct an training dataset, which contains 723 images totally. These images are initially indexed and correspond to their origin article's theme. Then, we select some text paragraphs from the original article and edit them by various ways. After that, we store the edited text paragraphs as images to construct an query image dataset, which is used to evaluate the accuracy of the proposed approach in various situations.

3.2 Experimental Results and Analyses

In the experimental procedure, we use the MatConvNet, which is a MATLAB toolbox, to train the CNN feature and mine the similarity of text images. The text database is composed of text paragraph images that converted from 6 English and 6 Chinese academic papers.

In the first case study, we choose a English text image that converted from the abstract of the fifth English article with various modifications, including retranslating by Google, changing the font color, adding another statement in the text, omitting lots of content, adjusting the line spacing of the text and reversing the word order. The query text image is shown as Figure 2(a). After that, we calculate the query text image's similarity to each text image in dataset by using article theme relevance filtering, the original text image that can be retrieved first is shown in Figure 2(b).

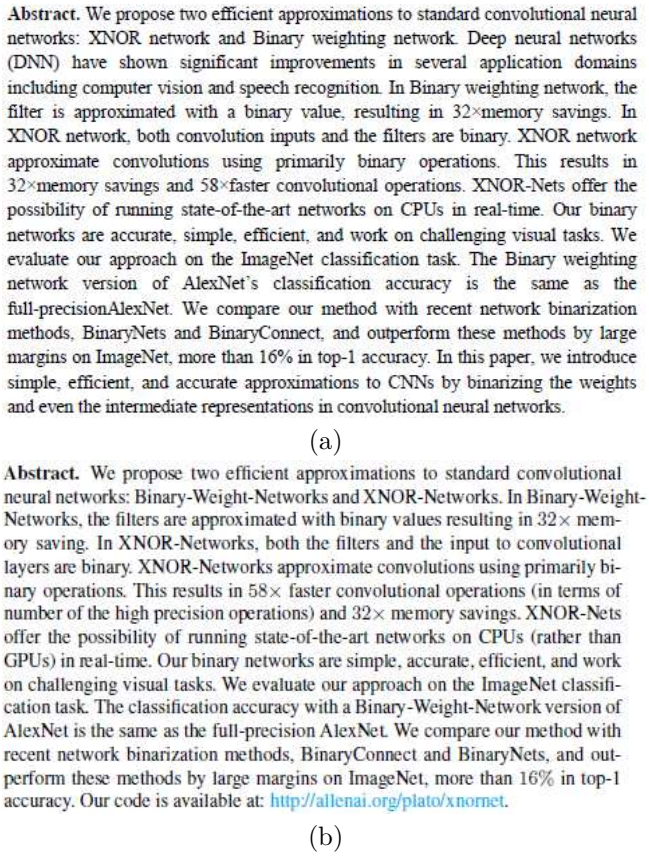

Fig. 2. The similarity retrieval of English text image. (a) The English query text image. (b) The original content retrieved from text image dataset.

Table 2. The article theme similarity of the text image as shown in Figure 2

Article theme	N=10	N=20
T1	0.9258	0.9075
T2	0	0.8909
T3	0.9257	0.9142
T4	0	0
T5	**0.9497**	**0.9297**
T6	0.9122	0.9059

Furthermore, we analyze the cosine similarity in the preliminary results of the English query text image similarity retrieval, and obtain the relevance of the query text image to each theme when the number of retrieval results N is 10 and 20. As shown in Table 2, T1–T6 stand for the themes of six different English articles respectively. In Table 2, similarity 0 means that there're no images affiliated to the article in the the preliminary results. The article theme relevance filtering model can be used to update the content similarity and optimize the retrieval accuracy, it can be seen that the query text image is most semantically similar to T5. Noted that there are no Chinese article retrieved via the query text image, so the article theme similarities of Chinese articles are all 0 and not listed in the table.

Another case study is provided to display the effect of the text similarity retrieval with theme relevance filtering when querying a Chinese text image that is re-translated by Google and modified in its original content. The retrieval result is shown in Figure 3, in which we can see that different visual presentations with similar text content can be found out by the proposed approach.

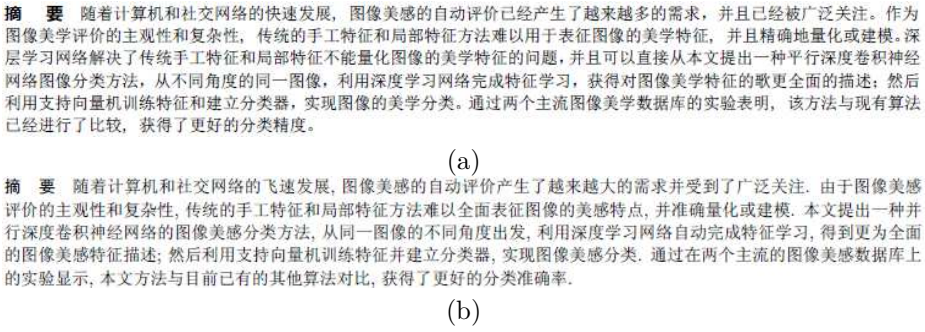

Fig. 3. The similarity retrieval of Chinese text image. (a) The Chinese query text image. (b) The original content retrieved from text image dataset.

Table 3. The similarity and correct order in different conditions.

Condition	English		Chinese	
	Similarity	Correct order	Similarity	Correct order
C1	0.9351	1	0.9601	1
C2	0.9424	1	0.9429	1
C3	0.8743	3	0.8299	2
C4	0.9266	2	0.9422	1
C5	0.9303	3	0.9123	1
C6	0.9437	1	0.9182	2
C7	0.9031	1	0.9002	2
C8	0.9279	2	0.9000	2

Furthermore, we use the similarity and the order of text image corresponded to the original content named correct order to measure the similarity retrieval system's quality. We respectively select a English text paragraph and Chinese one, and modify them in various ways to test the system's performance. In order to evaluate the similarity retrieval effect of the system on various types of text variation, the similarity and correct order of these experiments are shown in Table 3, where the number of retrieval results $N = 10$. In the table, condition C1 represents the experiment text image have no change basically with its original content, C2 represents adjusting the character spacing and line spacing, C3 means that there is a lot of blanks around the text image, C4 means adding another statement in the text, C5 stands for re-translating most of the words and sentence structure by Google, C6 represents increasing the font color changes based on C5, C7 is a complex condition including re-translating by Google, changing the font color, adding another statement in the text, omitting some content and adjusting the line spacing of the text, condition C8 is as the same as in Figure 2.

The above results represent that the similarity retrieval using CNN features can reflect the visual characteristics of text image. In different conditions listed in Table 3, the original content can be retrieved via a text image with correct order less than 3, which shows relatively high accuracy. Nevertheless, the approach may well be worth improving for higher retrieval accuracy and better adaptivity in more complex conditions.

4 Conclusions

In this paper, a new learning-based approach to text image retrieval is proposed by using CNN features extracted from text image and improved similarity metrics. The deep visual representations are initialized by a pre-trained network model and retrained for target task with PCA compression, which improves the adaptivity and increase retrieval accuracy of the CNN significantly. Besides that,

a theme relevance filtering model is proposed and integrated into the proposed approach to improve the retrieval accuracy. Experimental results show that good performance can be obtained for retrieving the original text content via a query text image. When this approach is further improved to adapt more complex text transformations, it is expected to be applied in paper plagiarism identification or literature recommendation.

References

1. Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. arXiv preprint arXiv:1505.07376, 12.
2. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).
3. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-9).
4. Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.
5. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
6. Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., & Shet, V. (2013). Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082.
7. Vedaldi, A., & Lenc, K. (2015, October). Matconvnet: Convolutional neural networks for matlab. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 689-692). ACM.
8. Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531.
9. Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014, September). Neural codes for image retrieval. In European conference on computer vision (pp. 584-599). Springer International Publishing.
10. Sejal, D., Ganeshsingh, T., Venugopal, K. R., Iyengar, S. S., & Patnaik, L. M. (2016). Image Recommendation Based on ANOVA Cosine Similarity. Procedia Computer Science, 89, 562-567.
11. Sejal, D., Rashmi, V., Venugopal, K. R., Iyengar, S. S., & Patnaik, L. M. (2016). Image recommendation based on keyword relevance using absorbing Markov chain and image features. International Journal of Multimedia Information Retrieval, 5(3), 185-199.